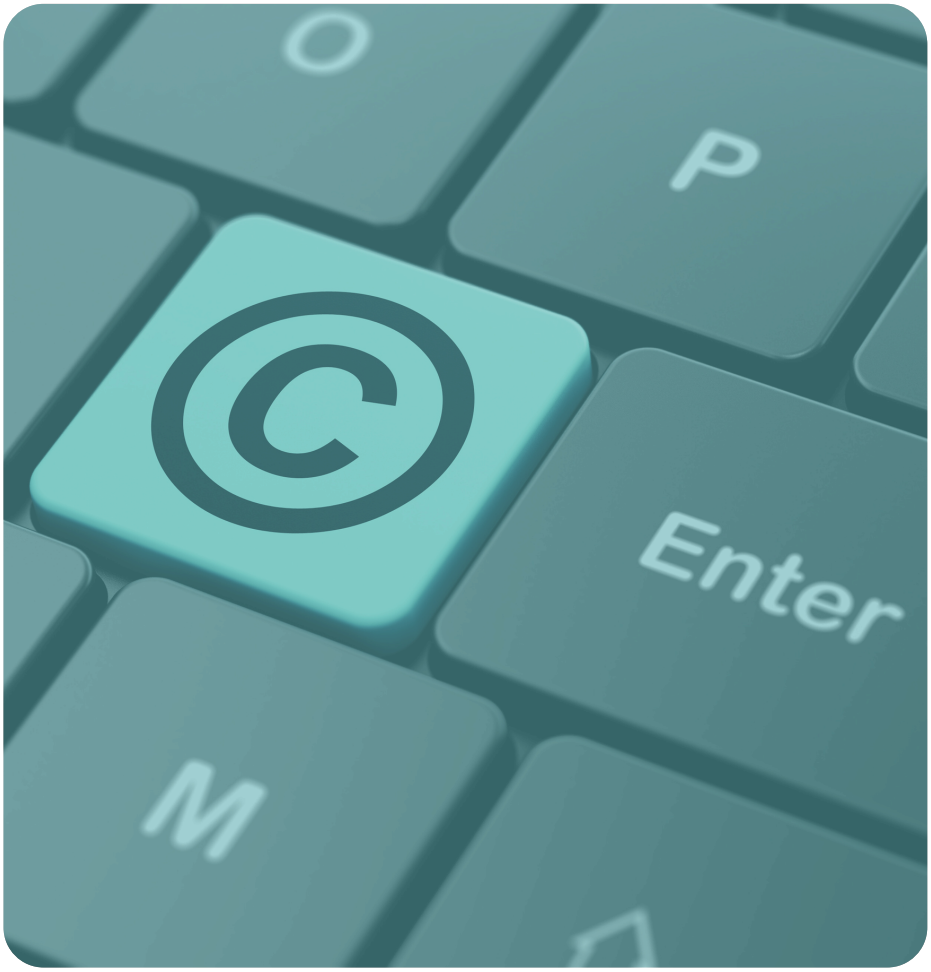
PRADO VIDIGAL

IA e Direitos Autorais: Caminhos para um Marco Legal Equilibrado



Novembro 2025 - v2.0

Índice

Clique para navegar

1. Desmistificando o Treinamento de Modelos de IA Generativa

A ascensão da Inteligência Artificial (IA) generativa representa um dos mais significativos desafios tecnológicos e intelectuais para o ordenamento jurídico contemporâneo. Legisladores em todo o mundo, inclusive no Brasil, debruçam-se sobre a tarefa de regular uma tecnologia cujo funcionamento interno é, muitas vezes, contraintuitivo e mal compreendido. O debate em torno do Projeto de Lei nº 2.338/2023 (PL 2338/2023) é um exemplo paradigmático desse desafio. Esforços legislativos, embora bem-intencionados na proteção de direitos fundamentais como os direitos autorais, correm o risco de, se não forem ancorados em uma sólida compreensão técnica, estabelecer entraves jurídicos, negociais e estruturais ao desenvolvimento da IA no país, resultando em barreira intransponível à inovação.

A raiz de muitas propostas regulatórias problemáticas reside em um profundo descompasso com a realidade técnica dos modelos de linguagem de grande escala (LLMs), os motores por trás de sistemas generativos como o tão popular ChatGPT. Conceitos jurídicos tradicionais de "cópia", "obra" e "autoria" são frequentemente aplicados de forma inadequada a um processo que não é de reprodução, mas de abstração e síntese estatística. O resultado é um atrito regulatório que não se origina de um conflito irreconciliável entre inovação e direitos, mas de uma lacuna de entendimentos entre o campo jurídico e o computacional. A lei tenta regular o que percebe como um ato de apropriação expressiva, enquanto a tecnologia opera em um plano de análise matemática de padrões que muito destoa daquilo que se entende tradicionalmente como violação de direitos (autorais ou de personalidade).

Para o jurista, o legislador e o formulador de políticas públicas, a ambientação técnica em IA deixou de ser um conhecimento acessório para se tornar uma competência essencial[1]. A formulação de leis eficazes, que fomentem a inovação ao mesmo tempo em que mitigam riscos reais, depende criticamente da superação de analogias falhas e da adoção de um modelo mental que reflita com precisão como esses sistemas aprendem e operam.

---

[1] UNIDERP. Inteligência Artificial para advogados: usos e desafios. Blog da Uniderp, 25 ago. 2025. Disponível em: https://blog.uniderp.com.br/inteligencia-artificial-para-advogados-usos-e-desafios/. Acesso em: 19 out. 2025.

1.1 Como funciona o treinamento?

Para adequada compreensão de como os modelos de IA generativa são construídos, é válido utilizar a analogia (não perfeita) da comparação do processo de treinamento de um LLM à jornada de aprendizado de um ser humano, como um jurista que se forma ao longo de anos de estudo. Essa comparação serve como um andaime conceitual para entender o processo como instrumento de absorção, abstração e síntese, afastando desde o início a premissa errônea de que a IA funciona como uma "colagem" ou "remixagem" de seus dados de treinamento.

A tabela a seguir estabelece, de maneira superficial mas didática, os paralelos entre o aprendizado humano e o de máquina:

Etapa	Aprendizagem de um humano (ex: jurista)	Treinamento de LLM (motor da IA generativa)	Considerações técnico-jurídicas
Exposição a dados (input)	Leitura de vasta doutrina, jurisprudência, legislação e artigos	Processamento de um massivo corpus de dados textuais (livros, artigos, websites etc.)	O humano lê para compreender e fruir da expressão. O LLM é exposto a textos como dados estatísticos para identificar padrões
Conhecimento	Memória sináptica no cérebro; conhecimento integrado e difuso, sem cópia literal do conteúdo lido	O "conhecimento" é codificado nos valores numéricos de bilhões de parâmetros (pesos e vieses) da rede neural	O cérebro não armazena cópias. O LLM, de forma análoga, não armazena cópias dos dados de treino; ele ajusta seus parâmetros
Aplicação / Criação (output)	Redige uma petição, parecer ou artigo original, aplicando os princípios aprendidos a um novo problema	Gera um novo texto, prevendo a sequência de *tokens* mais provável em resposta a um *prompt*, com base nos padrões estatísticos aprendidos	A criação humana é guiada pela intenção e consciência. A geração do LLM é uma síntese probabilística. O processo para chegar a uma saída não é de cópia, e sim de geração.

O equívoco comum é presumir que o treinamento de uma IA é um ato equivalente a "colagem" ou "remixagem" de obras protegidas. Na realidade, o processo é uma cascata de transformações em favor da análise estatística.

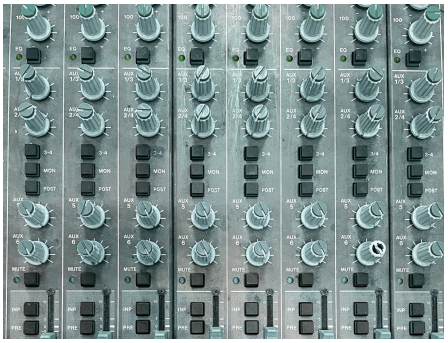
Primeiramente, o texto original é fragmentado em "tokens", que são pequenas unidades computacionais. A tokenização é o processo de segmentar uma sequência de texto em unidades menores, chamadas tokens. Imagine ensinar uma criança a ler: em vez de apresentar um parágrafo complexo de uma só vez, o processo começa com letras individuais, depois sílabas e, finalmente, palavras inteiras. De forma análoga, a tokenização divide o texto em pedaços digeríveis para a máquina, permitindo que os algoritmos identifiquem padrões com mais facilidade[2].

Um token pode ser uma palavra inteira ("advogado"), uma parte de uma palavra ou um prefixo ("in-" em "inconstitucional"), um caractere individual ou um sinal de pontuação. Em seguida, esses tokens são traduzidos para a linguagem da matemática através de "embeddings" (imersão vetorial). Cada token é convertido em um vetor (uma lista de números) que representa sua posição em um "mapa de conceitos" multidimensional[3]. O significado é capturado por relações geométricas, de modo que, por exemplo, o vetor para "juiz" estará próximo de "tribunal", mas longe de "fotossíntese". A expressão original é substituída por uma abstração matemática e o conteúdo é convertido em bilhões de representações vetoriais sem vínculos diretos com o conteúdo de origem.

Uma vez que os dados textuais foram desconstruídos e traduzidos para a linguagem da matemática, o processo de aprendizado pode começar[4]. Esse aprendizado ocorre dentro de uma estrutura computacional complexa inspirada no cérebro humano: a rede neural artificial. Para os LLMs, a arquitetura predominante é a chamada Transformer, introduzida em 2017, que revolucionou o campo por sua eficiência em processar longas sequências de texto[5]. No entanto, mais importante do que a arquitetura específica é entender como o "conhecimento" é armazenado. Diferente de um banco de dados convencional, um LLM não possui uma área de armazenamento para os textos que leu. Em vez disso, todo o seu aprendizado é codificado na configuração de seus componentes internos: os parâmetros (pesos e vieses).

---

[2] SMARTIA Solutions. O que é embeddings? Entenda sua importância. Blog da Smartia Solutions. Disponível em: https://smartiasolutions.com.br/glossario/o-que-e-embeddings-entenda-sua-importancia/. Acesso em: 20 out. 2025.
[3] GOOGLE DEVELOPERS. Embeddings: espaço de embedding e embeddings estáticos: Google for Developers, 16 maio 2025. Disponível em: https://developers.google.com/machine-learning/crash-course/embeddings/embedding-space?hl=pt-br. Acesso em: 20 out. 2025.
[4] ELASTIC. O que são embeddings vetoriais? | Um guia abrangente de embeddings vetoriais. Disponível em: https://www.elastic.co/pt/what-is/vector-embedding. Acesso em: 20 out. 2025.
[5] RUSH, Alexander; NGUYEN, Vincent; KLEIN, Guillaume. The Annotated Transformer. Harvard NLP, 3 abr. 2018. Disponível em: https://nlp.seas.harvard.edu/2018/04/03/attention.html . Acesso em: 20 out. 2025.



A analogia mais útil aqui é a de um sistema de som com bilhões de "botões" de ajuste (os parâmetros) [6]. O processo de treinamento consiste em girar e ajustar meticulosamente cada um desses botões até que o sistema consiga captar o "sinal" (os padrões subjacentes da linguagem humana) de forma clara, a partir do "ruído" (o imenso e caótico conjunto de dados de treinamento).

Ao final do treinamento, o modelo de IA é, fundamentalmente, o conjunto final desses valores numéricos. O produto entregue é a configuração otimizada de todos os seus parâmetros. O conhecimento do modelo sobre gramática, fatos, estilos de escrita e relações conceituais está inteiramente codificado, de forma distribuída e abstrata, nesses valores numéricos. Isso invalida tecnicamente a ideia de que o modelo é um "contêiner" ou uma "colagem" de dados. A informação foi transformada em uma função matemática complexa, parametrizada por esses valores.

Portanto, um LLM não armazena cópias do conteúdo ao qual teve acesso. O processo é de generalização, que consiste em aprender as regras e padrões subjacentes da linguagem, e não de compressão ou memorização. O objetivo central do treinamento de um LLM fundamental é surpreendentemente simples: aprender a prever o próximo token em uma sequência. Por ser resultado de um processo de generalização, o modelo tem um incentivo estatístico para "ignorar" ou "esquecer" características únicas em favor de padrões comuns.

Por outro lado, fato é que há casos em que acontece a “memorização” (ou “regurgitação”). Porém, é preciso considerar que a memorização não é o objetivo do treinamento; é um efeito colateral indesejado, uma falha no processo de generalização[7]. Ela é análoga ao overfitting em modelos de aprendizado de máquina mais simples, onde o modelo aprende os dados de treinamento "bem demais", incluindo o ruído e as particularidades [8], e perde sua capacidade de generalizar para novos dados. É fundamental notar que a comunidade de pesquisa em IA considera a memorização um problema a ser resolvido, não uma funcionalidade a ser explorada[9].

---

[6] DAG, Drew. Introduction to neural networks — weights, biases and activation. Medium, 27 Mar. 2023. Disponível em: https://medium.com/@theDrewDag/introduction-to-neural-networks-weights-biases-and-activation-270ebf2545aa. Acesso em: 20 out. 2025.

[7] MORRIS, John X.; SITAWARIN, Chawin; GUO, Chuan; et al. How much do language models memorize? arXiv, May 30, 2025. Disponível em: https://arxiv.org/html/2505.24832v1 . Acesso em: 20 out. 2025.

[8] SAKIB, Md Nazmus; ISLAM, Md Athikul; PATHAK, Royal; et al. Risks, Causes, and Mitigations of Widespread Deployments of Large Language Models (LLMs): A Survey. arXiv preprint arXiv:2408.04643v1, 2024. Disponível em: https://arxiv.org/html/2408.04643v1 . Acesso em: 20 out. 2025

[9] HANS, Abhimanyu; KIRCHENBAUER, John; WEN, Yuxin; JAIN, Neel; KAZEMI, Hamid; SINGHANIA, Prajwal; SINGH, Siddharth; SOMEPALLI, Gowthami; GEIPING, Jonas; BHATELE, Abhinav; GOLDSTEIN, Tom. Be like a Goldfish, Don’t Memorize! Mitigating Memorization in Generative LLMs.. Disponível em: https://neurips.cc/virtual/2024/poster/96066. Acesso em: 20 out. 2025.

Sendo assim, do ponto de vista regulatório, tratar a memorização como a função primária do modelo seria um erro categórico. Seria o equivalente a proibir a fabricação de automóveis porque alguns modelos apresentam defeitos de freio. A abordagem jurídica correta é distinguir entre a operação pretendida e a falha ocasional.

A regulação da Inteligência Artificial, especialmente no que tange à sua intersecção com os direitos autorais, exige que o legislador transcenda analogias simplistas e fundamente suas decisões em uma compreensão precisa dos mecanismos em jogo. A formulação de políticas públicas eficazes depende da internalização de quatro conclusões técnicas fundamentais:

- O treinamento é um processo de **abstração matemática**, não de "reprodução".
- O conhecimento do modelo **reside em seus parâmetros**, não em cópias de conteúdo.
- A geração de conteúdo é **probabilística**, não uma "colagem".
- Eventual memorização é vista como **falha**, e não a função pretendida.

Com base nessas premissas, torna-se claro que um arcabouço regulatório eficaz deve evitar a imposição de obrigações inexequíveis baseadas em modelos mentais equivocados, sobretudo porque o treinamento de modelos de IA é fundamental para o avanço da sociedade, especialmente do campo da IA Responsável.

1.2 Por que o treinamento não pode ser visto como vilão?

A adoção da IA em larga escala representa uma das maiores oportunidades de crescimento econômico para o Brasil nas próximas décadas. Análises de diversas instituições de renome convergem para um cenário de impacto substancial no Produto Interno Bruto (PIB) do país, condicionado a um ambiente regulatório que fomente a inovação.

+13 pontos no PIB

Estudos da consultoria PwC projetam que a IA tem o potencial de adicionar até 13 pontos percentuais ao PIB do Brasil ao longo da década de 2030[10], um incremento comparável ao processo de industrialização do século XIX.

+4,2% de crescimento

Em um horizonte mais curto, um estudo encomendado pela Microsoft à FrontierView estima que, em um cenário otimista de adoção, a IA poderia adicionar 4,2 pontos percentuais de crescimento adicional ao PIB brasileiro até 2030[11].

Esses ganhos macroeconômicos não são automáticos; são impulsionados fundamentalmente por um salto na produtividade. A IA permite a automação de tarefas repetitivas, a otimização de processos complexos e a geração de insights a partir de grandes volumes de dados, permitindo decisões mais rápidas e precisas. O FMI estima que a adoção de IA possa elevar a Produtividade Total dos Fatores (PTF) do Brasil em 1% a 4%, um indicador relevante da eficiência econômica de um país[12].

É fundamental compreender a cadeia causal que conecta a regulação ao crescimento. As projeções de crescimento do PIB não são um destino inevitável, mas uma oportunidade condicionada a um ambiente que permita o desenvolvimento e a adoção da tecnologia. O desenvolvimento de sistemas de IA eficazes e seguros depende, por sua vez, do treinamento de modelos com vastos conjuntos de dados. Uma legislação que, como o PL 2338/2023 em sua forma atual, cria barreiras praticamente intransponíveis para o treinamento de modelos com fins comerciais, está, na prática, bloqueando o caminho para que o Brasil realize esse potencial econômico. A proposta de lei, se não for alterada, se tornará a causa direta de uma oportunidade histórica perdida.

Importante reconhecer que o impacto da IA transcende os agregados macroeconômicos, materializando-se em avanços concretos em setores vitais para o desenvolvimento e o bem-estar da população brasileira. A capacidade de treinar e aplicar modelos de IA localmente é o que permitirá ao Brasil resolver seus próprios desafios com ferramentas adaptadas à sua realidade.

---

[10] FIEE — Feira Internacional da Indústria Elétrica, Eletrônica, Energia, Automação e Conectividade. Entenda como a IA pode fazer o PIB do Brasil crescer 13 pontos percentuais na próxima década. Blog Negócios e Desenvolvimento, São Paulo, 3 abr. 2025. Disponível em: https://www.fiee.com.br/pt-br/blog/negocios-e-desenvolvimento/entenda-como-a-ia-pode-fazer-o-pib-do-brasil-crescer-13-pontos-p.html.

[11] MICROSOFT NEWS CENTER. A adoção de Inteligência Artificial pode adicionar 4,2 pontos percentuais de crescimento adicional ao PIB do Brasil até 2030, 8 dez. 2020. Disponível em: https://news.microsoft.com/pt-br/a-adocao-de-inteligencia-artificial-pode-adicionar-42-pontos-percentuais-de-crescimento-adicional-ao-pib-do-brasil-ate-2030/. Acesso em: 20 out. 2025.

[12] VELASCO, Fernando. Efeitos da inteligência artificial sobre o emprego e produtividade no Brasil. Blog do IBRE — FGV, 07 ago 2024. Disponível em: https://portal.fgv.br/artigos/efeitos-inteligencia-artificial-sobre-emprego-e-produtividade-brasil. Acesso em: 20 out. 2025.

Saúde:

O setor de saúde brasileiro já incorpora IA de forma acelerada, com 62,5% das instituições de saúde já utilizando a tecnologia em 2023[13]. Os casos de uso demonstram um potencial transformador:

- **Gestão e Eficiência**: hospitais utilizam IA para otimizar a gestão de leitos e recursos, melhorando o fluxo de pacientes e reduzindo custos operacionais[14].

- **Diagnóstico de Precisão**: laboratórios aplicam IA para analisar exames de imagem e dados genômicos com maior velocidade e precisão, impactando mensalmente entre 10 mil e 140 mil pacientes[15]. Cerca de 17% dos médicos brasileiros já utilizam IA generativa em suas rotinas, principalmente para suporte à pesquisa e elaboração de relatórios[16].

- **Desenvolvimento Farmacêutico**: a IA acelera a descoberta de novos medicamentos ao analisar interações moleculares e permite a personalização de tratamentos com base em perfis genômicos individuais[17].

- **Ampliação do Acesso**: a IA pode ajudar a mitigar as desigualdades regionais de acesso à saúde, conectando pacientes em áreas remotas a especialistas por meio de soluções de telemedicina e otimizando o fluxo de informações entre unidades de saúde[18].

---

[13] Conforme índices apresentados por: Associação Brasileira de Startups de Saúde (ABSS) & Associação Nacional de Hospitais Privados (ANHP). “62,5% das instituições de saúde no Brasil já usam IA”, diz CEO. Future Health, 16 maio 2025. Disponível em: https://futurehealth.cc/instituicoes-saude-incorporam-ia-inteligencia-artificial/ . Acesso em: 20 out. 2025.
[14] PWC. Casos públicos de uso de IA no setor de saúde: transformação tecnológica já impacta o presente e define o futuro da saúde. São Paulo: PwC Brasil, 10 abr. 2025. Disponível em: https://www.pwc.com.br/pt/estudos/setores-atividade/saude/2025/casos-publicos-de-uso-de-ia-no-setor-de-saude.html. Acesso em: 20 out. 2025.
[15] Ib Idem.
[16] SBIS, Sociedade Brasileira de Informática em Saúde. Estudo revela que 17% dos médicos no Brasil utilizam Inteligência Artificial, o número entre os enfermeiros é de 16%. SBIS, 21 jan. 2025. Disponível em: https://sbis.org.br/noticia/estudo-revela-que-17-dos-medicos-no-brasil-utilizam-inteligencia-artificial-o-numero-entre-os-enfermeiros-e-de-16/. Acesso em: 20 out. 2025.
[17] PWC. Casos públicos de uso de IA no setor de saúde: transformação tecnológica já impacta o presente e define o futuro da saúde. São Paulo: PwC Brasil, 10 abr. 2025. Disponível em: https://www.pwc.com.br/pt/estudos/setores-atividade/saude/2025/casos-publicos-de-uso-de-ia-no-setor-de-saude.html . Acesso em: 20 out. 2025.
[18] MEARIM, Marcelo. IA Generativa no setor da saúde no Brasil – Como alcançar o ROI e transformar a experiência do paciente. Saúde Business, 6 fev. 2025. Disponível em: https://www.saudebusiness.com/ti-e-inovao/ia-generativa-no-setor-da-saude-no-brasil/. Acesso em: 20 out. 2025.

Agronegócio

Em um setor que representa aproximadamente 27% do PIB brasileiro, a IA é a espinha dorsal da chamada "Agricultura 5.0", que transforma dados em decisões acionáveis no campo[19].

- **Agricultura de Precisão:** robôs autônomos, guiados por satélite e equipados com visão computacional, identificam e aplicam herbicidas apenas sobre plantas daninhas, reduzindo drasticamente o uso de defensivos e o impacto ambiental[20].

- **Logística Preditiva:** cooperativas de grãos utilizam IA para otimizar rotas de transporte, analisando em tempo real variáveis como previsões climáticas, custos de combustível e congestionamentos portuários, resultando em reduções de até 18% nos atrasos de entrega[21].

- **Gestão de Sustentabilidade (ESG):** a IA permite o monitoramento automático e em tempo real de indicadores de sustentabilidade, como emissões de carbono e consumo de água, cruzando dados de satélites, sensores de campo e documentos fiscais para gerar relatórios de conformidade de forma ágil e precisa[22].

Educação

A IA oferece a promessa de personalizar o ensino em uma escala sem precedentes, abordando desafios históricos do sistema educacional brasileiro.

- **Aprendizagem Adaptativa:** plataformas baseadas em IA avaliam o desempenho do aluno em tempo real e ajustam o conteúdo e a dificuldade das atividades para atender às suas necessidades individuais, permitindo que cada estudante avance em seu próprio ritmo[23].

---

[19] SERRAVALLE, Flávio. Agro 5.0: como a inteligência artificial está transformando o campo. Veja, 28 abr. 2024. Atualizado em 3 jun. 2024. Disponível em: https://veja.abril.com.br/economia/rumo-ao-agro-5-0-como-inteligencia-artificial-esta-transformando-o-campo/. Acesso em: 27 out. 2025.
[20] Ib. Idem.
[21] SOMATIVA. 5 exemplos de como a IA está revolucionando o agronegócio em todos seus departamentos. Blog Somativa, 17 jul. 2025. Disponível em: https://www.somativa.com.br/blog/5-exemplos-de-como-a-ia-esta-revolucionando-o-agronegocio-em-todos-seus-departamentos/. Acesso em: 27 out. 2025.
[22] Ib. Idem.
[23] GIANNINI, Fernando. 43 exemplos de inteligência artificial aplicados na educação. Blog Fernando Giannini. Disponível em: https://fernandogiannini.com.br/43-exemplos-de-inteligencia-artificial-aplicados-na-educacao/. Acesso em: 20 out. 2025.

- **Inclusão e Acessibilidade:** tecnologias assistivas, como softwares de reconhecimento de voz que transcrevem aulas para alunos com deficiência auditiva, promovem um ambiente de aprendizado mais inclusivo[24].

- **Otimização do Trabalho Docente:** a automação de tarefas administrativas e de correção libera tempo para que os professores se concentrem em atividades de maior valor, como o acompanhamento individualizado dos alunos e o planejamento pedagógico criativo[25].

A capacidade de treinar modelos de IA é a infraestrutura basilar da nova economia digital. Não se trata de uma atividade secundária, mas do núcleo do processo de criação de valor.

Conforme alertado por entidades como Brasscom, ABES e Abria, impor barreiras regulatórias a esta etapa, como a exigência de licenciamento prévio para cada dado utilizado em treinamentos com finalidade comercial, na prática, inviabiliza o desenvolvimento de IA no Brasil[26].

Esta barreira tem duas consequências diretas e danosas. A primeira é a perda de competitividade econômica. Em um mercado global que movimenta trilhões de dólares, o Brasil seria relegado à condição de mero consumidor de tecnologia, importando modelos desenvolvidos e treinados no exterior.

A segunda consequência é a perda de soberania digital. Modelos treinados no exterior, naturalmente, apresentam dificuldades em compreender as nuances do português brasileiro, a diversidade cultural do país ou os problemas específicos de nossa sociedade, o que é agravado caso a regulação imponha barreiras excessivas aos dados necessários ao treinamento. Uma IA treinada com dados predominantemente norte-americanos ou europeus, por exemplo, será menos eficaz para diagnosticar doenças prevalentes na população brasileira, para otimizar a logística do agronegócio nacional ou para personalizar a educação de acordo com as diretrizes curriculares do Brasil.

---

[24] Ib. Idem.
[25] CASSOL, Daniela. Quais os impactos do ChatGPT e da Inteligência Artificial na Educação. Portal IFSC, publicado em 2 ago. 2022. Disponível em: https://www.ifsc.edu.br/web/ifsc-verifica/w/quais-os-impactos-do-chatgpt-e-da-inteligencia-artificial-na-educacao- . Acesso em: 20 out. 2025.
[26] BRASSCOM. Posicionamento do Setor Produtivo por regras equilibradas de direitos autorais e pela competitividade do Brasil no cenário global de inteligência artificial. Brasília: Brasscom, 3 dez. 2024. Disponível em: https://brasscom.org.br/pdfs/por-regras-equilibradas-de-direitos-autorais-e-pela-competitividade-do-brasil-no-cenario-global-de-inteligencia-artificial/. Acesso em: 20 out. 2025.

Restringir o treinamento local é, portanto, uma renúncia à capacidade de construir soluções tecnológicas soberanas e verdadeiramente adaptadas às necessidades do país.

Sistemas de Inteligência Artificial aprendem a partir dos dados com os quais são alimentados. Eles não possuem uma compreensão inata de conceitos como justiça ou equidade; eles identificam e replicam os padrões presentes em seus dados de treinamento. Consequentemente, se os dados refletem os vieses, preconceitos e desigualdades históricas da sociedade, o modelo de IA resultante não apenas reproduzirá, mas frequentemente amplificará essas distorções de forma sistêmica e em larga escala.

A principal e mais eficaz estratégia técnica para mitigar o viés algorítmico é garantir que os conjuntos de dados de treinamento sejam o mais amplos, diversos e representativos possível da população do mundo real[27]. Dados limitados ou não representativos são a causa raiz de múltiplos tipos de vieses prejudiciais, conforme detalhado na tabela abaixo[28]:

Tipo de viés	Descrição	Exemplo	Possível solução
Viés de Seleção	Ocorre quando os dados de treinamento não são representativos da população que o modelo encontrará no mundo real.	Sistemas de reconhecimento facial que apresentam taxas de erro significativamente maiores para mulheres e pessoas de pele negra.	Inclusão massiva de imagens que representem toda a diversidade de tons de pele, gêneros e etnias, garantindo que o modelo aprenda a identificar características faciais de forma equitativa.

[27] IBM. O que é viés de dados? IBM Think. Disponível em: https://www.ibm.com/br-pt/think/topics/data-bias. Acesso em: 20 out. 2025.
[28] CHAPMAN UNIVERSITY. Bias in AI. Chapman University, 2025. Disponível em: https://www.chapman.edu/ai/bias-in-ai.aspx . Acesso em: 20 out. 2025.

Tipo de viés	Descrição	Exemplo	Possível solução
Viés de Confir-mação / histórico	Ocorre quando o modelo aprende e reforça preconceitos e desigualdades presentes em dados históricos.	O algoritmo de recrutamento que penalizava currículos contendo termos associados a mulheres, pois foi treinado com base em contratações passadas, majoritariamente masculinas.	Aumento do conjunto de dados para incluir exemplos de sucesso de grupos sub-representados e aplicação de técnicas de rebalanceamento para que o modelo não associe sucesso a um único perfil demográfico.
Viés de Estereoti-pagem	Ocorre quando o modelo associa certas características ou papéis a grupos específicos, reforçando estereótipos sociais.	Modelos que, quando solicitado a retratar uma pessoa advogada bem-sucedida, criará a imagem de um homem, hétero, branco, de meia-idade, com cabelos grisalhos e vestindo roupas sociais de luxo.	Exposição do modelo a um volume de textos e imagens muito maior e mais variado, incluindo literatura, artigos acadêmicos e notícias que mostrem pessoas de todos os gêneros em diversas profissões, corrigindo as associações estereotipadas.
Viés de Homoge-neidade do Grupo Externo	Ocorre quando o modelo generaliza indivíduos de grupos sub-representados, tratando-os como mais similares entre si do que realmente são.	Sistemas de IA que têm dificuldade em diferenciar indivíduos de minorias étnicas, levando a classificações incorretas e potenciais falsas identificações em contextos de segurança.	Coleta e inclusão de um número estatisticamente significativo de dados de cada grupo demográfico, permitindo que o modelo aprenda a distinguir indivíduos com a mesma acurácia para todos.

A governança de IA e a gestão de vieses são, portanto, processos técnicos que dependem fundamentalmente da matéria-prima utilizada: os dados. Sem acesso a dados em quantidade e variedade suficientes, qualquer esforço para construir uma IA justa está fadado ao fracasso[29].

Impor barreiras legais e econômicas ao acesso a dados para treinamento no Brasil, como um regime de licenciamento prévio obrigatório (atingindo inclusive os dados publicamente disponíveis), criaria um cenário paradoxal e perigoso. Em vez de proteger os cidadãos, tal medida os exporia a riscos sistêmicos de discriminação algorítmica. Os desenvolvedores nacionais, impedidos de treinar seus modelos com dados que refletem a vasta diversidade brasileira (racial, regional, de gênero, socioeconômica), seriam forçados a um de dois caminhos indesejáveis:

1. Importação de Vieses Estrangeiros: o mercado brasileiro seria inundado por modelos de IA desenvolvidos no exterior, predominantemente treinados com dados do Norte Global. Esses sistemas chegariam ao Brasil carregando preconceitos alheios à nossa realidade e incapazes de compreender as especificidades da nossa população, resultando em discriminação e ineficácia.

2. Criação de Sistemas "Ingênuos" e Ineficazes: os poucos modelos desenvolvidos localmente seriam treinados com conjuntos de dados limitados e não representativos. Tais sistemas seriam frágeis, incapazes de lidar com a complexidade do mundo real e propensos a gerar resultados imprecisos e discriminatórios ao se depararem com situações não previstas em seu treinamento limitado.

Fato é que uma IA mal treinada pode ser um rápido caminho para a perpetuação de ciclos de desigualdade. A causa não está no excesso de dados, mas justamente na sua escassez e falta de diversidade/qualidade. Restringir o acesso a informações para treinamento é, paradoxalmente, garantir que os sistemas reproduzam apenas as perspectivas dominantes, ignorando vozes e contextos marginalizados. Se queremos IAs justas, precisamos de bases de dados amplas, diversas e representativas, e não de muros que limitem o que essas tecnologias podem aprender sobre o mundo real.

---

[29] Sobre o tema, recomendamos pesquisa elaborada pelo Centro de Ensino e Pesquisa em Inovação (CEPI) da FGV Direito SP, conforme: SILVA, Alexandre P. da; FEFERBAUM, Marina; COSTA, Enya C. S. da; ZAVAGLIA COELHO, Alexandre; MORALES, Luiza X.; NOMURA, M. M.; RIBEIRO, Maria F. F.. Governança da Inteligência Artificial em Organizações: Arquitetura de Confiabilidade e Gestão de Vieses. São Paulo: CEPI FGV Direito SP, 2025. Disponível em: <https://repositorio.fgv.br/server/api/core/bitstreams/bebe8525-90d7-49ac-a6ae-c9e3aa3ddd79/content>.

2. Desafios da Atual Redação do PL 2.338/2023

Uma vez considerados os aspectos técnicos do treinamento de modelos de IA, bem como sua importância para que tenhamos soluções úteis, eficazes e justas, o debate em torno da regulação e IA deve assumir mais duas premissas centrais: (i) a de que Direitos Autorais são direitos fundamentais, tal como expressamente previsto em nossa Constituição Federal e, portanto, não podem ter sua proteção desconsiderada por conta do avanço tecnológico e (ii) é papel do Direito encontrar a ponderação ideal entre proteção a direitos e inovação, por mais desafiadora que a tarefa pareça ser. Nossa análise parte, portanto, do profundo respeito aos direitos autorais, buscando identificar pontos de aprimoramento na proposta atual que permitam sua melhor compatibilização com o inegável avanço tecnológico.

Isso posto, fato é que, diante das massivas e notórias críticas que os atuais artigos 62 a 65 do PL 2338/2023 vêm sofrendo, parece que a principal proposta legislativa que temos atualmente em debate no Brasil ainda não logrou êxito em encontrar esse necessário ponto de equilíbrio.

A "Seção IV – Dos Direitos de Autor e Conexos" do PL 2.338/2023, embora tenha o mérito de buscar a criação de um regime jurídico que previna eventuais lesões aos Direitos Autorais, acaba por estabelecer entraves que podem se mostrar intransponíveis ao desenvolvimento de sistemas de inteligência artificial no Brasil, afastando o Brasil de outras jurisdições que vêm estruturando seus marcos regulatórios sobre IA de forma mais equilibrada e factível.

2.1 Obrigatoriedade de publicação de sumário (art. 62)

O artigo 62 estabelece que os desenvolvedores de sistemas de inteligência artificial devem publicar um sumário dos conteúdos protegidos utilizados nas etapas de mineração, treinamento, retreinamento, testagem, validação e aplicação, preservando segredos comerciais e industriais:

> *Art. 62. O desenvolvedor de IA que utilizar conteúdo protegido por direitos de autor e conexos deverá informar sobre os conteúdos protegidos utilizados nos processos de desenvolvimento dos sistemas de IA, por meio da publicação de sumário em sítio eletrônico de fácil acesso, observados os segredos comercial e industrial, nos termos de regulamento específico.*
> *Parágrafo único. Para fins deste Capítulo, o desenvolvimento compreende as etapas de mineração, treinamento, retreinamento, testagem, validação e aplicação de sistemas de IA.*

A intenção de garantir transparência, embora pareça legítima à primeira vista, gera problemas operacionais em virtude da redação proposta, revelando um descompasso entre obrigação regulatória e a realidade técnica dos LLMs.

Como visto no capítulo anterior, o treinamento de modelos de larga escala, como os modelos de linguagem, envolve o processamento de trilhões de tokens e de múltiplas fontes de dados[30]. Esses conteúdos são convertidos em representações estatísticas dispersas nos parâmetros do modelo, tornando tecnicamente inviável realizar uma identificação, ou rastrear, de maneira verificável, quais obras específicas contribuíram para os resultados extraídos pelo uso de um sistema.

Exigir a divulgação de um "sumário" de conteúdos protegidos implica uma expectativa tecnicamente inviável e potencialmente violadora de segredos industriais[31].

Mesmo que se admitisse a possibilidade de listar classes ou conjuntos de dados, o nível de detalhamento necessário para que essa divulgação tivesse valor jurídico seria de tal ordem que acabaria por expor práticas e diferenciais empresariais.

Tanto é que, conforme emenda (não acolhida) apresentada pelo Sen. Carlos Portinho durante tramitação da proposta de lei no Senado Federal[32], o "volume massivo de dados" utilizados por sistemas de IA torna "excessivamente difícil" a identificação de obras utilizadas, razão pela qual o dispositivo, em sua forma atual, carece de proporcionalidade e razoabilidade para sua dimensão prática.

---

[30] Sobre o funcionamento e operacionalização dos sistemas, entende Bertin Martens: "GenAI foundation models learn by extracting insights from a huge pool of human knowledge and skills that resides in books, documents and texts, webpages, art and audio-visual products, etc. The sheer scale of that pool and algorithmic learning, using trillions of word tokens as inputs, results in unexpected emerging learning properties, such as the ability to learn in-context and apply insights to categories of data that were not part of the pool. GenAI models are still on an exponential growth path in terms of data pooling and computing capacity". MARTENS, Bertin. Economic arguments in favour of reducing copyright protection for generative AI inputs and outputs. Working Paper 09/2024: Bruegel, 2024, p. 12.

[31] Destacamos que essa preocupação não é unívoca do cenário brasileiro. Nesse sentido, Boniface de Champris escreve: "From a business perspective, however, these new provisions raise several questions and will likely make it more difficult for providers of foundation models to innovate. Indeed, the type and amount of content used to train such models can make a substantial difference between the performance of one model or another. In many instances, the content will be proprietary to the model developer and the relevant sources used to train a given model can well constitute business-confidential information or trade secrets. In such a dynamic and competitive market, the public disclosure of the content used to train a given foundation model can be particularly sensitive and needs to be subject to strong safeguards in line with Union and national law". CHAMPRIS, Boniface de. Artificial Intelligence and Copyright: The EU Should Preserve the Copyright Directive's Delicate Balance to Safeguard and Promote Innovation. AIRe – Journal of AI Law and Regulation, v.1, 2024, p. 116.

[32] A emenda apresentada pode ser encontrada em: <https://legis.senado.leg.br/sdleg-getter/documento?disposition=inline&dm=9853989>. Acesso em 15 de out. 2025.

2.2 Restrição à mineração de textos e dados (art. 63)

O artigo 63, por sua vez, cria uma exceção para a mineração automatizada de textos e dados (*TDM - Text and Data Mining*), limitando-a a instituições científicas, educacionais, museus, arquivos públicos e bibliotecas, desde que sem fins comerciais. Além disso, proíbe expressamente o envolvimento de entidades vinculadas ou coligadas a empresas com fins lucrativos:

> *Art. 63. Não constitui ofensa aos direitos de autor e conexos a utilização automatizada de conteúdos protegidos em processos de mineração de textos e dados para os fins de pesquisa e desenvolvimento de sistemas de IA por organizações e instituições científicas, de pesquisa e educacionais, museus, arquivos públicos e bibliotecas, desde que observadas as seguintes condições:*
> *I – o acesso tenha se dado de forma lícita;*
> *II – não tenha fins comerciais;*
> *III – a utilização de conteúdos protegidos por direitos de autor e conexos seja feita na medida necessária para o objetivo a ser alcançado, sem prejuízo dos interesses econômicos dos titulares e sem concorrência com a exploração normal das obras e conteúdos protegidos.*
> *§ 1º Cópias de conteúdos protegidos por direitos de autor e conexos utilizadas nos sistemas de IA deverão ser armazenadas em condições de segurança, e unicamente pelo tempo necessário para a realização da atividade ou para a finalidade específica de verificação dos resultados.*
> *§ 2º É vedada a exibição ou a disseminação das obras e conteúdos protegidos por direitos de autor e conexos utilizados no desenvolvimento de sistemas de IA.*
> *§ 3º Este artigo não se aplica a instituições vinculadas, coligadas ou controladas por entidade com fins lucrativos que forneçam sistemas de IA ou que tenham, entre elas, participação acionária.*
> *§ 4º Aplica-se o disposto no caput deste artigo à mineração de dados, por entidades públicas ou privadas, no contexto de sistemas de IA para combate a ilícitos civis e criminais, que atentem contra direitos de autor e conexos.*

Esse desenho normativo cria um regime de exceção demasiadamente restritivo, desalinhado com as práticas internacionais e com a própria lógica contemporânea da pesquisa científica e tecnológica. A bem da verdade, a redação atual do artigo 63 cria um paradoxo: ao mesmo tempo em que o legislador parece reconhecer a importância da mineração de textos e dados para avanços na ciência, educação e cultura por meio da IA, ele restringe severamente a possibilidade de que tais campos sejam desenvolvidos com o apoio da iniciativa privada – que hoje tem participação fundamental em todos eles.

Como exposto anteriormente, a mineração de dados necessária para treinamento de modelos de IA não utiliza obras para fruição expressiva, mas as analisa como dados. O objetivo não é fruir da obra, mas analisá-la computacionalmente para extrair padrões.

Nesse sentido, o reconhecimento da legalidade do input não imuniza o output de ser considerado infrator a depender do caso em concreto[33].

Além disso, o modelo regulatório proposto ignora as melhores práticas internacionais e se afasta das políticas de países que são referência em inovação:

Jurisdição	Permissão para TDM?	Mecanismo Chave
União Europeia	**Sim.**	Exceção regulatória (Diretiva UE 2019/790, Arts. 3 e 4).
Estados Unidos	**Sim** (analisada caso a caso).	Doutrina do *Fair Use* (Uso Justo), conforme interpretação do 17 U.S.C. § 107.
Japão	**Sim.**	Exceção legal ampla (Copyright Act, Art. 30-4).
Singapura	**Sim.**	Doutrina do *Fair Use* (Uso Justo), conforme interpretação do Copyright Act.
Israel	**Sim** (mediante interpretação).	Parecer do Ministério da Justiça[34] (não vinculante) estabeleceu que TDM para ML é geralmente *Fair Use*, conforme interpretação do Copyright Act.

Fato é que, a adoção do artigo 63 da forma como está poderá criar uma estrutura regulatória isolada e pouco competitiva a nível global. Em outras palavras, a consequência prática dessa restrição é a criação de uma barreira intransponível para o desenvolvimento de IA no Brasil, beneficiando empresas estrangeiras que operam sob outras jurisdições mais amigáveis ao avanço tecnológico.

---

[33] GARDNER, Stephen. Anthropic Copyright Ruling: What It Means for AI and Fair Use. Berger Singerman, 15 jul. 2024. Disponível em: https://www.bergersingerman.com/news-insights/anthropic-copyright-ruling-what-it-means-for-ai-and-fair-use. Acesso em 16 out. 2025.

[34] Disponível em: https://www.gov.il/BlobFolder/legalinfo/machine-learning/he/machine-learning.pdf. Acessado em 20 out. 2025.

O resultado prático será, potencialmente, o agravamento da sub-representação do conteúdo em língua portuguesa nos modelos globais. Modelos treinados com bases de dados que limitam o conhecimento sobre a sociedade brasileira podem levar à perpetuação de preconceitos, injustiças históricas e sub-representação cultural brasileira, perdendo a livre-iniciativa nacional a oportunidade de treinar modelos com materiais que "reflitam suas próprias características, línguas, demandas e desejos"[35].

Em suma, a redação do artigo 63, que prevê a exceção para a mineração de textos e dados limitada a organizações de determinados setores e sem fins comerciais, é um entrave ao desenvolvimento tecnológico, social e econômico. Em vez de equilibrar direitos, o dispositivo em questão propicia um ambiente de insegurança jurídica que punirá a inovação brasileira e desestimulará o treinamento de modelos – etapa essencial do desenvolvimento tecnológico responsável.

2.3 Proibição/*opt-out* (art. 64)

O artigo 64 é confuso em sua essência. Enquanto o artigo 63 inviabiliza, como regra geral, a mineração de texto e dados, o artigo 64 confere o direito de os titulares de direitos autorais proibirem a utilização de seus conteúdos no desenvolvimento de sistema de IA (mecanismo globalmente conhecido como opt-out):

> *Art. 64. O titular de direitos de autor e conexos poderá proibir a utilização dos conteúdos de sua titularidade no desenvolvimento de sistemas de IA nas hipóteses não contempladas pelo art. 63 desta Lei.*
> *Parágrafo único. A proibição do uso de obras e conteúdos protegidos nas bases de dados de um sistema de IA posterior ao processo de treinamento não exime o agente de IA de responder por perdas e danos morais e materiais, nos termos da legislação aplicável.*

Ocorre que o mecanismo do opt-out apenas faz sentido quando a legislação autoriza o processamento de obras protegidas para determinada finalidade, como, por exemplo, para fins de mineração e treinamento de modelos de IA. Ao limitar a possibilidade de mineração a hipóteses mínimas e ainda estabelecer o direito de proibição, o PL 2.338/2023 institui um regime de dupla restrição que parece se afastar do ponto de equilíbrio entre proteção a direitos autorais e fomento ao desenvolvimento tecnológico.

---

[35] Conforme entendem Luca Schirru, Allan Rocha, Mariana Valente e Alice Lana: “On a cultural and social level, countries will lose an important opportunity to train models with materials that reflect their own characteristics, languages, demands and desires in different fields, including but not limited to health (e.g. neglected diseases) and culture (e.g. linguistic peculiarities)”. SCHIRRU, Luca; SOUZA, Allan Rocha de; VALENTE, Mariana Giorgetti; LANA, Ana de Paula. Text and Data Mining Exceptions in Latin America. IIC – International Review of Intellectual Property and Competition Law, v. 55, 2024, p. 1648.

Diferentemente do modelo adotado pela União Europeia, cujo artigo 4 da Diretiva 2019/790[36] estabelece uma exceção geral para mineração de textos e dados com acesso lícito, o legislador brasileiro, segundo leitura conjugada dos artigos 63 e 64, parece ter adotado até o momento o opt-in (autorização prévia) como regra inclusive para mineração de textos e dados publicamente disponíveis.

Ao fazê-lo, o legislador do PL 2.338/2023 expande em demasia a própria proteção legal inerente aos direitos autorais, que não serve para restringir acesso nem mesmo extração de conhecimento de conteúdo disponibilizado/acessado de maneira pública e/ou lícita.

Vale destacar que hoje já existem mecanismos técnicos amplamente difundidos, como (i) o uso do arquivo robots.txt, padrão consolidado para restringir o acesso automatizado a websites; (ii) a incorporação de metadados e marcações semânticas, a exemplo dos modelos do Creative Commons e Schema.org, que permitem incluir licenças e restrições diretamente no conteúdo; (iii) soluções de machine readable opt-out, empregadas em plataformas como arXiv, GitHub e jornais digitais para indicar a possibilidade ou não de indexação e reutilização de dados; e (iv) barreiras técnicas como paywalls, logins obrigatórios e autenticações por cookies, amplamente utilizadas por editoras e veículos de mídia para impedir a mineração automatizada não autorizada, caso assim desejem fazer os detentores de direitos autorais.

O resultado prático das restrições impostas pelo PL 2.338/2023 é que, em última instância, elas tornam ilícita a extração de conhecimento a partir de textos, áudios e imagens legalmente acessados.

Analogamente, o que a proposta legislativa faz é tornar ilícita a extração de conhecimento de um livro aberto, disponível a qualquer um que o queira ler. Para além de desafiar os limites da proteção autoral, as restrições pretendidas pela atual versão da proposta legislativa em comento geram um evidente efeito paralisante sobre o avanço tecnológico no Brasil, com alto risco de fuga de capitais intelectuais e tecnológicos para ambientes regulatórios mais equilibrados.

---

[36] "Artigo 4º - Exceções ou limitações para a prospeção de textos e dados
1. Os Estados-Membros devem prever uma exceção ou uma limitação aos direitos previstos no artigo 5.o, alínea a), e no artigo 7.o, n.o 1, da Diretiva 96/9/CE, no artigo 2.o da Diretiva 2001/29/CE, no artigo 4.o, n.o 1, alíneas a) e b), da Diretiva 2009/24/CE e no artigo 15.o, n.o 1, da presente diretiva, para as reproduções e as extrações de obras e de outro material protegido legalmente acessíveis para fins de prospeção de textos e dados.
2. As reproduções e extrações efetuadas nos termos do n.o 1 podem ser conservadas enquanto for necessário para fins de prospeção de textos e dados.
3. A exceção ou limitação prevista no n.o 1 é aplicável desde que a utilização de obras e de outro material protegido a que se refere esse número não tenha sido expressamente reservada pelos respetivos titulares de direitos de forma adequada, em particular por meio de leitura ótica no caso de conteúdos disponibilizados ao público em linha.
4. O presente artigo não prejudica a aplicação do artigo 3.o da presente diretiva".

2.4 Remuneração obrigatória pela utilização de conteúdos protegidos (art. 65)

Ratificando a arquitetura de sobreposição de restrições de acesso a dados para treinamento de IA, o artigo 65 determina que os titulares de direitos autorais sejam remunerados pela utilização de seus conteúdos em processos de mineração, treinamento e desenvolvimento de sistemas de inteligência artificial, ainda que o acesso aos dados tenha sido legítimo.

> *Art. 65. O agente de IA que utilizar conteúdos protegidos por direitos de autor e conexos em processos de mineração, treinamento ou desenvolvimento de sistemas de IA deve remunerar os titulares desses conteúdos em virtude dessa utilização, devendo-se assegurar:*
> *I – que os titulares de direitos de autor e conexos tenham condições efetivas de negociar coletivamente, nos termos do Título VI da Lei nº 9.610, de 19 de fevereiro de 1998 (Lei dos Direitos Autorais), ou diretamente a utilização dos conteúdos dos quais são titulares, podendo fazê-lo de forma gratuita ou onerosa;*
> *II – que o cálculo da remuneração a que se refere o caput considere os princípios da razoabilidade e da proporcionalidade e elementos relevantes, tais como o porte do agente de IA e os efeitos concorrenciais dos resultados em relação aos conteúdos originais utilizados;*
> *III – a livre negociação na utilização dos conteúdos protegidos, visando à promoção de ambiente de pesquisa e experimentação que possibilite o desenvolvimento de práticas inovadoras, e que não restrinjam a liberdade de pactuação entre as partes envolvidas, nos termos dos arts. 156, 157, 421, 422, 478 e 479 da Lei nº 10.406, de 10 de janeiro de 2002 (Código Civil), e o art. 4º da Lei nº 9.610, de 19 de fevereiro de 1998 (Lei dos Direitos Autorais).*
> *§ 1º A remuneração a que se refere o caput deste artigo é devida somente:*
> *I – aos titulares de direitos de autor e conexos nacionais ou estrangeiros domiciliados no Brasil;*
> *II – a pessoas domiciliadas em país que assegure a reciprocidade na proteção, em termos equivalentes a este artigo, aos direitos de autor e conexos de brasileiros, conforme disposto nos arts. 2º, parágrafo único, e 97, § 4º, da Lei nº 9.610, de 19 de fevereiro de 1998 (Lei dos Direitos Autorais), sendo vedada a cobrança nos casos em que a reciprocidade não estiver assegurada.*
> *§ 2º O titular do direito de remuneração previsto no caput que optar pela negociação e autorização direta, nos termos do inciso I do caput, poderá exercê-lo independentemente de regulamentação posterior.*

Para além das considerações já feitas nos itens anteriores, o modelo de remuneração obrigatória sobre a “utilização” de conteúdo publicamente disponível trazido pelo PL 2.338/2023 carece de viabilidade operacional. O treinamento de um modelo de linguagem não é um ato de cópia, mas um processo de transformação fundamental que converte dados brutos em representações matemáticas abstratas. O conteúdo original é desconstruído em tokens e transformado em um gigantesco modelo de pesos estatísticos que, quando muito, encapsula padrões linguísticos, não as obras literais.

Nesse processo, o vínculo causal com qualquer obra específica é efetivamente rompido, tornando a tarefa calcular uma remuneração individual proporcional algo que, no atual estado da técnica, beira o impossível[37]. Exigir tal feito seria o equivalente a pedir que um humano listasse e remunerasse o autor de cada frase que já leu e que contribuiu para formar seu estilo de escrita[38].

Ao insistir em um modelo de remuneração obrigatória para o input, o Brasil se afasta drasticamente das abordagens mais equilibradas e pró-inovação adotadas por outras jurisdições.

Nos Estados Unidos, embora o tema esteja longe de um consenso, o treinamento de IA começa a ser enquadrado na doutrina do fair use como um uso "transformativo". A jurisprudência americana, a exemplo do caso Bartz et al. v. Anthropic PBC[39], considerou o treinamento de um modelo de linguagem com base em livros "espetacularmente transformador".

Sob a ótica jurídico-constitucional, o artigo 65, como desenhado, parece não ser capaz de superar o teste constitucional de proporcionalidade[40]. A medida soa inadequada, pois cobrar pelo input não se mostra apto a endereçar o dano relevante, que emerge apenas no resultado. É potencialmente desnecessária, pois existem meios menos gravosos e mais eficazes, como mecanismos de opt-out técnico. E, por fim, poderia ser considerada desproporcional, pois os custos sociais e o efeito paralisante sobre o desenvolvimento tecnológico e a propagação de conhecimento parecem superar os eventuais benefícios.

Ademais, vale reiterar que as regras de direitos autorais excessivamente rigorosas dispostas na proposta legislativa afastam o Brasil de ter alguma relevância na geopolítica competitiva da IA. Em potencial, a norma seria facilmente executável contra empresas brasileiras, que estarão sujeitas à fiscalização e a sanções em território nacional, mas de aplicação prática bastante limitada contra desenvolvedores estrangeiros baseados em jurisdições mais flexíveis.

Na prática, o Brasil estaria legislando contra si mesmo, impondo um custo e um risco à inovação nacional que seus concorrentes internacionais não terão.

---

[37] RAMOS, Pedro Henrique; BARRETO, Julia de Albuquerque; GARROTE, Marina. Remuneração por Direitos Autorais em IA: Limites e Desafios de Implementação. Série Policy Briefs. São Paulo: Reglab, 2025. n. 3, p. 04.
[38] SOBEL, Benjamin. Artificial Intelligence's Fair Use Crisis. 41 Colum. J.L. & Arts 45, 2017, pp. 57 e 58.
[39] Mais em: Mais informações sobre a decisão podem ser encontradas em: <https://www.afslaw.com/perspectives/alerts/landmark-ruling-ai-copyright-fair-use-vs-infringement-bartz-v-anthropic>. Acesso em 16 out. 2025.
[40] Cf. SILVA, Luís Virgílio Afonso da. O proporcional e o razoável. Revista dos Tribunais, São Paulo, v.91, n.798, 2002, p. 34; e ÁVILA, Humberto. A distinção entre princípios e regras e a redefinição do dever de proporcionalidade. Revista Diálogo Jurídico, Salvador, v. 1, n. 4, 2001.

Nesse sentido, as consequências para o empreendedorismo local seriam profundamente prejudiciais, pois é inimaginável que tal camada da sociedade brasileira consiga negociar valores absorvíveis por seus negócios em razão do "uso" de informação publicamente disponível para gerar aprendizados[41].

Portanto, para as startups os custos de transação e a incerteza jurídica seriam proibitivos, funcionando como uma barreira de entrada intransponível que inibe a inovação local e sufoca a concorrência. O resultado não seria apenas a regulação das grandes (que talvez tenham poder econômico para adquirir bases de dados licenciadas[42]), mas a eliminação da capacidade de inovação das pequenas e médias empresas brasileiras que dependem do processamento de dados publicamente disponíveis.

3. Sugestões concretas para o PL 2.338/2023

Diante das considerações técnicas e jurídicas anteriormente expostas, bem como das duras críticas que a Seção IV do PL 2.338/2023 vêm sofrendo, parece-nos claro que deve haver outro caminho que melhor harmonize a inafastável proteção aos Direitos Autorais com o acesso à informação que se faz vital para o desenvolvimento de soluções de IA que sejam bem recebidas pela sociedade.

Com isso, reiterando nosso respeito ao direito fundamental da proteção autoral, apresentamos a seguir possíveis caminhos que tentam enfrentar a difícil missão de conciliar valores nessa esfera do debate regulatório. O objetivo não é anular nem reduzir o direito autoral, mas interpretá-lo de forma a equilibrar a proteção dos criadores com o imperativo do progresso tecnológico e do acesso ao conhecimento – pilares que justificam o próprio monopólio garantido pela via da proteção autoral. A inovação em IA e a proteção autoral não são, e não devem ser, objetivos mutuamente excludentes, e é com esse espírito que apresentamos as alternativas a seguir, sendo que, a nosso ver, todas elas equilibram melhor a proteção autoral com o necessário acesso a dados para evolução da IA.

---

[41] "While licensing possibilities are indeed evolving and AI developers are entering into agreements with high-value content providers (e.g., news organizations and stock photography companies), the report underestimates the significant economic burdens of widespread licensing requirements for AI training, particularly for smaller companies, researchers, and the open-source community. Sweeping licensing requirements would stifle innovation in AI development, fostering a monopolistic environment dominated by large technology companies while encouraging opportunistic litigation and raising costs for consumers". BROUGH, Wayne. Copyright, AI Training, and Innovation. R. Street, 2025. Disponível em: <https://www.rstreet.org/commentary/copyright-ai-training-and-innovation/>. Acesso em 16 out. 2025.

[42] Nesse sentido, explica Bertin Martens, em trabalho acerca dos argumentos econômicos acerca do treinamento de sistemas de IA: "Big GenAI firms with deep pockets are able to pay for license agreements. This may push smaller GenAI start-ups out of the market and reduce competition in GenAI services. It may also lead to biased selection of inputs for training of GenAI models". MARTENS, Bertin. Economic arguments in favour of reducing copyright protection for generative AI inputs and outputs. Working Paper 09/2024, Bruegel, 2024

3.1 Alteração no artigo 2º (Fundamentos)

O treinamento de IA é parte vital do desenvolvimento de sistemas seguros, eficazes e eticamente bem aceitos pela sociedade. No entanto, o atual texto do PL 2.338/2023 parece não reconhecer tal fato de forma adequada ao trazer muito mais entraves do que segurança ao treinamento de modelos de IA.

Proposta de Redação	Justificativa
Art. 2º, XXI - a promoção e legitimação do treinamento, retreinamento e aperfeiçoamento de sistemas de inteligência artificial como condição essencial para o desenvolvimento de sistemas seguros, transparentes, eficazes e não discriminatórios.	Consideramos importante que a promoção ao treinamento de IA seja positivada como fundamento da futura lei, de forma a trazer não apenas segurança jurídica à atividade como também servir de direcionamento geral aos entes regulados quanto à importância de tal etapa para o avanço tecnológico responsável.

3.2 Alterações nos artigos 62 a 65

Alternativa 1 **Exceção Ampliada para Mineração de Textos e Dados (TDM)**

Esta alternativa alinha o Brasil a jurisdições mais competitivas em termos de desenvolvimento de IA. Ela reconhece que o treinamento de IA a partir de dados acessados licitamente, para além de essencial para o desenvolvimento tecnológico, é processo que passa por um uso transformativo de conteúdo (e não reprodução indevida).

Proposta de Redação	Justificativa
Art. 62. (Removido)	A exigência de sumário é removida por ser tecnicamente inviável.

Proposta de Redação	Justificativa
Art. 63. Não constitui ofensa aos direitos de autor e conexos a utilização automatizada de conteúdos, textos e dados para fins de mineração, treinamento, retreinamento, testagem e validação de sistemas de inteligência artificial, desde que o acesso a tais conteúdos tenha se dado de forma lícita.	A nova redação generaliza a permissão de TDM, removendo a restrição "sem fins comerciais" que atualmente aniquila o desenvolvimento pela iniciativa privada. Isso fomenta o potencial econômico e permite o treinamento de modelos com dados diversos (desde que lícitos), que refletem a realidade brasileira e podem ser úteis inclusive para mitigação de vieses.
Art. 64. (Removido)	A remuneração obrigatória (Art. 65) e o opt-out (Art. 64) são removidos por serem incompatíveis com uma exceção legal ampla e por sua inviabilidade operacional.
Art. 65. (Removido)	

Alternativa 2 Exceção Ampliada para Mineração de Textos e Dados (TDM) com Opt-out

Esta alternativa, alinhada aos fundamentos deste documento, estabelece a mineração de dados como lícita (inclusive para fins comerciais) quando o acesso é lícito, contrabalanceando-a com um mecanismo de *opt-out* tecnicamente operável.

Proposta de Redação	Justificativa
Art. 62. (Removido)	A exigência de sumário é removida por ser tecnicamente inviável.

Proposta de Redação	Justificativa
Art. 63 - Mesma redação da alternativa 1.	Essa solução evita a completa inviabilidade técnica e econômica da proposta original. Ela permite o aprendizado (abstração matemática, não cópia), que é crucial para a inovação e mitigação de vieses, ao mesmo tempo em que permite a utilização de um opt-out tecnicamente operável (robots.txt etc.) por parte dos titulares de direitos autorais.
Art. 64 O titular de direitos de autor e conexos poderá opor-se à utilização prevista no art. 63, sem efeitos retroativos, por meio de mecanismos técnicos eficazes e legíveis por máquina que indiquem expressamente a reserva de direitos no ponto de acesso ao conteúdo.	
Art. 65. (Removido)	

Alternativa 3 **Supressão Integral da Seção**

Caso nenhuma das alternativas anteriores soe conciliadora o suficiente, restaria ao legislador a opção de remoção dos artigos 62 a 66 da proposta legislativa, partindo-se da premissa de que a matéria é estranha ao objeto principal do PL 2.338 (LCP 95, Art. 7º), sendo disciplinada por outras legislações específicas (como a Lei nº 9.610/98, no caso dos direitos autorais, e Código Civil, no caso dos direitos de personalidade).

4. Conclusão

O direito autoral, concebido para uma era analógica de cópias físicas, possui a flexibilidade intrínseca para se adaptar à revolução da Inteligência Artificial. Uma legislação moderna e equilibrada, alinhada com precedentes internacionais, permite enquadrar o treinamento de IA (o input) como um uso de conteúdo que é, ao mesmo tempo, transformador e legítimo.

Esse uso, por ser de natureza técnica e não expressiva, não afeta a exploração normal das obras legitimamente acessadas e ainda promove o avanço tecnológico responsável, um bem social maior.

Referências Bibliográficas

Associação Brasileira de Startups de Saúde (ABSS) & Associação Nacional de Hospitais Privados (ANHP). “62,5% das instituições de saúde no Brasil já usam IA”, diz CEO. Future Health, 16 maio 2025. Disponível em: https://futurehealth.cc/instituicoes-saude-incorporam-ia-inteligencia-artificial/. Acesso em: 20 out. 2025.

ÁVILA, Humberto. A distinção entre princípios e regras e a redefinição do dever de proporcionalidade. Revista Diálogo Jurídico, Salvador, v. 1, n. 4, 2001.

BRASSCOM. Posicionamento do Setor Produtivo por regras equilibradas de direitos autorais e pela competitividade do Brasil no cenário global de inteligência artificial. Brasília: Brasscom, 3 dez. 2024. Disponível em: https://brasscom.org.br/pdfs/por-regras-equilibradas-de-direitos-autorais-e-pela-competitividade-do-brasil-no-cenario-global-de-inteligencia-artificial/. Acesso em: 20 out. 2025.

BROUGH, Wayne. Copyright, AI Training, and Innovation. R. Street, 2025. Disponível em: https://www.rstreet.org/commentary/copyright-ai-training-and-innovation/. Acesso em: 16 out. 2025.

CASSOL, Daniela. Quais os impactos do ChatGPT e da Inteligência Artificial na Educação. Portal IFSC, 2 ago. 2022. Disponível em: https://www.ifsc.edu.br/web/ifsc-verifica/w/quais-os-impactos-do-chatgpt-e-da-inteligencia-artificial-na-educacao-. Acesso em: 20 out. 2025.

CHAMPRIS, Boniface de. Artificial Intelligence and Copyright: The EU Should Preserve the Copyright Directive’s Delicate Balance to Safeguard and Promote Innovation. AIRe – Journal of AI Law and Regulation, v. 1, 2024, p. 116.

CHAPMAN UNIVERSITY. Bias in AI. Chapman University, 2025. Disponível em: https://www.chapman.edu/ai/bias-in-ai.aspx. Acesso em: 20 out. 2025.
DAG, Drew. Introduction to neural networks — weights, biases and activation. Medium, 27 mar. 2023. Disponível em: https://medium.com/@theDrewDag/introduction-to-neural-networks-weights-biases-and-activation-270ebf2545aa. Acesso em: 20 out. 2025.

ELASTIC. O que são embeddings vetoriais? | Um guia abrangente de embeddings vetoriais. Disponível em: https://www.elastic.co/pt/what-is/vector-embedding. Acesso em: 20 out. 2025.

FIEE — Feira Internacional da Indústria Elétrica, Eletrônica, Energia, Automação e Conectividade. Entenda como a IA pode fazer o PIB do Brasil crescer 13 pontos percentuais na próxima década. Blog Negócios e Desenvolvimento, São Paulo, 3 abr. 2025. Disponível em: https://www.fiee.com.br/pt-br/blog/negocios-e-desenvolvimento/entenda-como-a-ia-pode-fazer-o-pib-do-brasil-crescer-13-pontos-p.html.

GARDNER, Stephen. Anthropic Copyright Ruling: What It Means for AI and Fair Use. Berger Singerman, 15 jul. 2024. Disponível em: https://www.bergersingerman.com/news-insights/anthropic-copyright-ruling-what-it-means-for-ai-and-fair-use. Acesso em: 16 out. 2025.

GIANNINI, Fernando. 43 exemplos de inteligência artificial aplicados na educação. Blog Fernando Giannini. Disponível em: https://fernandogiannini.com.br/43-exemplos-de-inteligencia-artificial-aplicados-na-educacao/. Acesso em: 20 out. 2025.

GOOGLE DEVELOPERS. Embeddings: espaço de embedding e embeddings estáticos. Google for Developers, 16 maio 2025. Disponível em: https://developers.google.com/machine-learning/crash-course/embeddings/embedding-space?hl=pt-br. Acesso em: 20 out. 2025.

HANS, Abhimanyu; KIRCHENBAUER, John; WEN, Yuxin; JAIN, Neel; KAZEMI, Hamid; SINGHANIA, Prajwal; SINGH, Siddharth; SOMEPALLI, Gowthami; GEIPING, Jonas; BHATELE, Abhinav; GOLDSTEIN, Tom. Be like a Goldfish, Don't Memorize! Mitigating Memorization in Generative LLMs. Disponível em: https://neurips.cc/virtual/2024/poster/96066. Acesso em: 20 out. 2025.

IBM. O que é viés de dados? IBM Think. Disponível em: https://www.ibm.com/br-pt/think/topics/data-bias. Acesso em: 20 out. 2025.

MARTENS, Bertin. Economic arguments in favour of reducing copyright protection for generative AI inputs and outputs. Working Paper 09/2024: Bruegel, 2024, p. 12.

MEARIM, Marcelo. IA Generativa no setor da saúde no Brasil – Como alcançar o ROI e transformar a experiência do paciente. Saúde Business, 6 fev. 2025. Disponível em: https://www.saudebusiness.com/ti-e-inovao/ia-generativa-no-setor-da-saude-no-brasil/. Acesso em: 20 out. 2025.

MICROSOFT NEWS CENTER. A adoção de Inteligência Artificial pode adicionar 4,2 pontos percentuais de crescimento adicional ao PIB do Brasil até 2030, 8 dez. 2020. Disponível em: https://news.microsoft.com/pt-br/a-adocao-de-inteligencia-artificial-pode-adicionar-42-pontos-percentuais-de-crescimento-adicional-ao-pib-do-brasil-ate-2030/. Acesso em: 20 out. 2025.

MORRIS, John X.; SITAWARIN, Chawin; GUO, Chuan; et al. How much do language models memorize? arXiv, 30 maio 2025. Disponível em: https://arxiv.org/html/2505.24832v1. Acesso em: 20 out. 2025.

PWC. Casos públicos de uso de IA no setor de saúde: transformação tecnológica já impacta o presente e define o futuro da saúde. São Paulo: PwC Brasil, 10 abr. 2025. Disponível em: https://www.pwc.com.br/pt/estudos/setores-atividade/saude/2025/casos-publicos-de-uso-de-ia-no-setor-de-saude.html. Acesso em: 20 out. 2025.

RAMOS, Pedro Henrique; BARRETO, Julia de Albuquerque; GARROTE, Marina. Remuneração por Direitos Autorais em IA: Limites e Desafios de Implementação. Série Policy Briefs, São Paulo: Reglab, 2025, n. 3, p. 04.

RUSH, Alexander; NGUYEN, Vincent; KLEIN, Guillaume. The Annotated Transformer. Harvard NLP, 3 abr. 2018. Disponível em: https://nlp.seas.harvard.edu/2018/04/03/attention.html. Acesso em: 20 out. 2025.

SAKIB, Md Nazmus; ISLAM, Md Athikul; PATHAK, Royal; et al. Risks, Causes, and Mitigations of Widespread Deployments of Large Language Models (LLMs): A Survey. arXiv preprint arXiv:2408.04643v1, 2024. Disponível em: https://arxiv.org/html/2408.04643v1. Acesso em: 20 out. 2025.

SBIS — Sociedade Brasileira de Informática em Saúde. Estudo revela que 17% dos médicos no Brasil utilizam Inteligência Artificial, o número entre os enfermeiros é de 16%. SBIS, 21 jan. 2025. Disponível em: https://sbis.org.br/noticia/estudo-revela-que-17-dos-medicos-no-brasil-utilizam-inteligencia-artificial-o-numero-entre-os-enfermeiros-e-de-16/. Acesso em: 20 out. 2025.

SCHIRRU, Luca; SOUZA, Allan Rocha de; VALENTE, Mariana Giorgetti; LANA, Ana de Paula. Text and Data Mining Exceptions in Latin America. IIC – International Review of Intellectual Property and Competition Law, v. 55, 2024, p. 1648.

SERRAVALLE, Flávio. Agro 5.0: como a inteligência artificial está transformando o campo. Veja, 28 abr. 2024. Atualizado em 3 jun. 2024. Disponível em: https://veja.abril.com.br/economia/rumo-ao-agro-5-0-como-inteligencia-artificial-esta-transformando-o-campo/. Acesso em: 27 out. 2025.

SILVA, Alexandre P. da; FEFERBAUM, Marina; COSTA, Enya C. S. da; ZAVAGLIA COELHO, Alexandre; MORALES, Luiza X.; NOMURA, M. M.; RIBEIRO, Maria F. F. Governança da Inteligência Artificial em Organizações: Arquitetura de Confiabilidade e Gestão de Vieses. São Paulo: CEPI FGV Direito SP, 2025. Disponível em: https://repositorio.fgv.br/server/api/core/bitstreams/bebe8525-90d7-49ac-a6ae-c9e3aa3ddd79/content.

SILVA, Luís Virgílio Afonso da. O proporcional e o razoável. Revista dos Tribunais, São Paulo, v. 91, n. 798, 2002, p. 34.

SMARTIA Solutions. O que é embeddings? Entenda sua importância. Blog da Smartia Solutions. Disponível em: https://smartiasolutions.com.br/glossario/o-que-e-embeddings-entenda-sua-importancia/. Acesso em: 20 out. 2025.
SOBEL, Benjamin. Artificial Intelligence's Fair Use Crisis. 41 Colum. J.L. & Arts 45, 2017, pp. 57 e 58.

SOMATIVA. 5 exemplos de como a IA está revolucionando o agronegócio em todos seus departamentos. Blog Somativa, 17 jul. 2025. Disponível em: https://www.somativa.com.br/blog/5-exemplos-de-como-a-ia-esta-revolucionando-o-agronegocio-em-todos-seus-departamentos/. Acesso em: 27 out. 2025.

UNIDERP. Inteligência Artificial para advogados: usos e desafios. Blog da Uniderp, 25 ago. 2025. Disponível em: https://blog.uniderp.com.br/inteligencia-artificial-para-advogados-usos-e-desafios/. Acesso em: 19 out. 2025.

VELASCO, Fernando. Efeitos da inteligência artificial sobre o emprego e produtividade no Brasil. Blog do IBRE — FGV, 7 ago. 2024. Disponível em: https://portal.fgv.br/artigos/efeitos-inteligencia-artificial-sobre-emprego-e-produtividade-brasil. Acesso em: 20 out. 2025.

Material elaborado por **Prado Vidigal Advogados** em 27.10.2025.

PRADO VIDIGAL